Team-13: Automatic Action Recognition

Team Members: Garrett Bartlow, Daniel Miao, Joshua Stadtmueller, Jonathan Zamudio, Braxton Parker

Sponsor: Dr. Khoa Luu

Problem Statement

- Rapidly growing advancements in the AI community have allowed us the ability to innovate the way users can present in PowerPoint like never before.

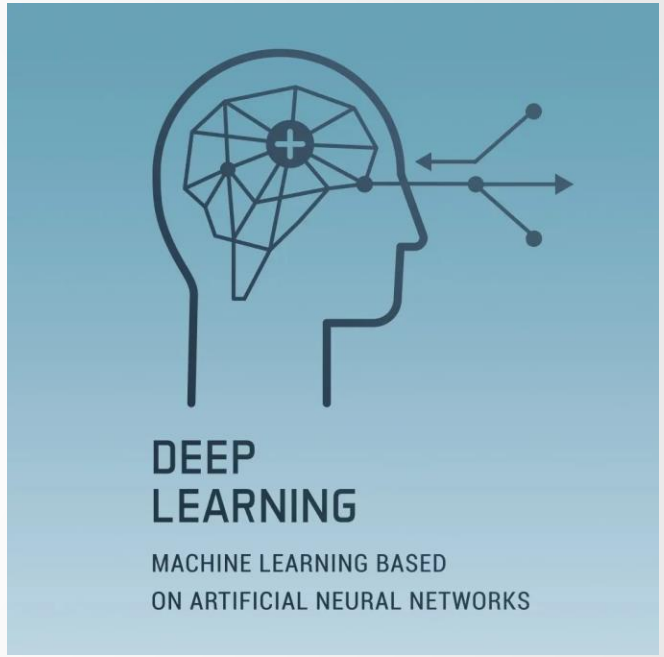
- **Automatic Action Recognition:** Expanding and implementing software that utilizes Automatic Action Recognition to allow users the ability to control presentations via hand gestures

Objectives

- Create a GUI using the Model-View-Controller method.
- Utilize deep learning libraries and video processing software to recognize gestures according to our standards.
- Design our project so that it may act as a launching platform for more complex variations in the future.

Background

- Deep Learning
  - "allow(s) computers to learn from experience and understanding the world in terms of a hierarchy of concepts"
  - Adept at solving abstract problems
  - Large amounts of data required
- Convolutional Neural Network (CNN)
  - Used mainly in the field of pattern recognition within images
  - Used on larger images
  - CNNs are comprised of neurons that self-optimize through learning



Background

- Temporal Shift Module (TSM)
  - Utilizes the temporal dimension of images to manipulate data to achieve 3D CNN results with 2D CNN complexity.
  - Takes in video, filters that video into collection of images, then predicts gesture in each image.

- Python, OpenCV, PyTorch, and Scikit-Learn
  - OpenCV: open-source library that deals with real-time computer vision
  - PyTorch and Scikit-Learn: machine learning libraries for Python that enable manipulation, construction, and analysis of data

Related Work

- MIT
  - Focus was increasing efficiency and implementing with Google Maps.
  - Our project's focus is to support a stand-alone application designed for an increased interfacing experience.
- ASL recognition
  - A team used CNN and glove with sensors, achieved 96% accuracy
  - Another team used infrared images fed into CNN, achieved 99.7% accuracy
  - Image/video data is better for CNNs
- Predict unknown gestures

These other works allow for inspiration and a proof-of-concept ideas for new projects. Video recognition can be paired with other technology and can lead to development of innovative technology.

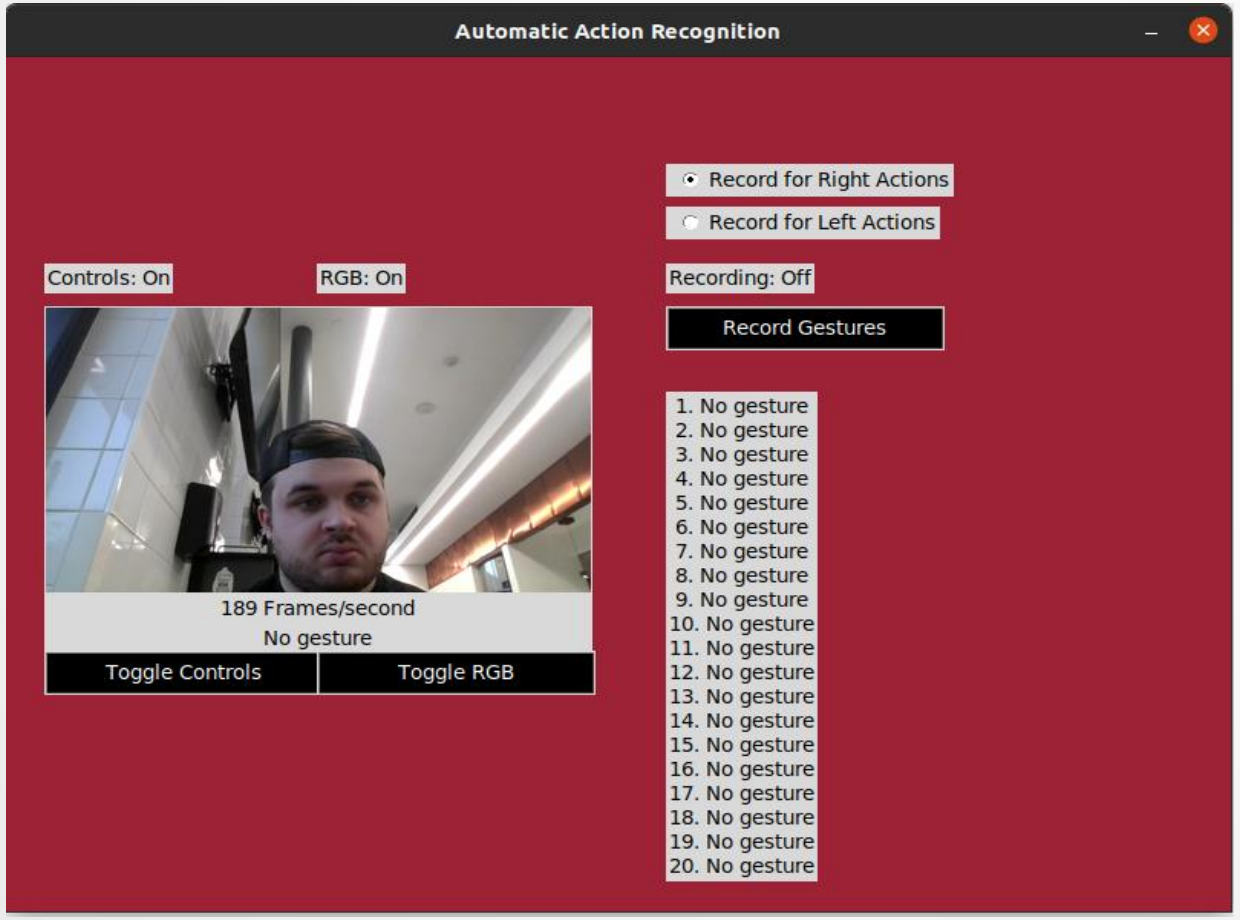
Architecture and Implementation

- Application consists of the "engine" and the user interface
- The engine consists of the model execution that is computed each program iteration
  - Engine takes in some webcam video stream and outputs the recognized gesture, if recognized (if not recognized, "No Gesture" is the output of the engine)
- User interface designed using the Model-View-Controller conventions (mostly)
  - Each view can be considered different if a component's state changes
    - Because of using Tkinter, no need for View class since states are easy to manage
    - No need to add complexity
- The user interface consists of components
  - Webcam
  - Buttons to:
    - Toggle computer controls
    - Toggle RGB for the webcam
    - Select which action is to be recorded
    - Record the selected action and update the gesture for that action

Implementation/Improvements

- For our specific application that is presenting by PowerPoint, we propose the following heuristic idea:
  - Some activities can inadvertently contain multiple gestures
  - "Calibrate" a particular activity as the sum or list of gestures done during that activity – done by the Record Gesture button
  - Assign that list to the chosen computer control
  - Specifically,
    - From the engine, if the current gesture has not been output the previous frame AND the previous gesture is not in the current recorded list, THEN execute the designated computer control/command
    - Assign the current gesture to the previous gesture such that the previous gesture at frame = frame + 1 is equivalent to the gesture of the current frame
- Retrained model on Something-Something V2, not a noticeable change except specific gestures

Interface Design



Lessons Learned and Impact

- Lessons Learned
  - A better understanding of the technologies used to create this application
  - Not adding un-necessary complexity to the project
  - Being able to create a hierarchy with related components

- Impact
  - Expand the capabilities to be able to interface with our technologies
  - Tangible impact: Being able to control PowerPoint presentations with simple gestures and no additional controllers

Future Work

- Improvements on the recognition models
  - Specialized model for each action, allowing for tighter restrictions on accepting a gesture
- Extend the accessibility of the project
  - Currently the application only works on Linux based systems
  - Better the user interface so that the application is easier to use\control